DOCUMENT RESUME

ED 282 418 FL 016 709

AUTHOR Samson, Digna M. M.
TITLE Rasch and Reading
PUB DATE May 83
NOTE 8p.; In: van Weeren, J., Ed. Practice and Problems in Language Testing 5; see FL 016 706. Small print in figures may affect legibility.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS English (Second Language); Foreign Countries; *Language Tests; *Multiple Choice Tests; *Reading Comprehension; Reading Skills; Reading Tests; Second Languages; Standardized Tests; *Statistical Analysis; Test Format; Testing Problems; *Test Items; *Test Validity
IDENTIFIERS Netherlands; *Rasch Model

ABSTRACT

The traditional multiple-choice reading comprehension test of English as a second language, used in the Dutch school-leaving examinations, has been criticized for its apparent lack of construct validity. The Dutch National Institute for Educational Measurement has conducted a number of studies to determine whether there is a different skill involved in answering multiple-choice questions compared to that used in summarizing the same reading passages. The results suggest that differential responses to the different test item types is more a function of the difficulty of the test item type than of the students' skills. However, the summary-type item appeared to measure skills more accurately, especially with higher-ability students. Further research on the use of the summary-type item is planned. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED282418

The traditional multiple choice reading comprehension test, used in the Dutch final examination, has always met with all sorts of criticisms. One of them is, "Our pupils have not really understood what the passages in the test are about, even though their scores on the multiple choice questions are satisfactory!"
We have tried to find out if there really is a different skill involved when pupils answer multiple choice questions in a reading comprehension test compared to when they summarize the reading passages used in the same test.

When the girl who was going to type the abstract for this paper asked me what title I wanted to give my paper, I perhaps too rashly thought of 'Rasch and Reading'. I am of course not going to deal with reading as a whole. Nor with reading in the mother tongue. What I do want to tell you a little about is about reading in a foreign language, or rather, about three ways of testing reading comprehension in a foreign language, and the analysis of the results of these tests. First, some background information. Here, at CITO, a lot of time and money is spent on the construction of multiple choice foreign language reading tests. They play an important role in the final examinations of secondary schools. Illustrative is the fact that a pupil's result on the test is weighted 50% in his examination mark. So there is plenty of reason to be very alert to criticism, and to try and improve the test. Over the years we have had quite a lot of it: most of it having to do with particular items or tests, or question types, but a considerable part of it was criticism on the method we used. So one important area needing attention is that of content validity. We have to see to it that the test measures the skill it is intended to measure. This is more easily said than done, of course. We do not have explicitly defined objectives for the required levels of reading comprehension. We only know that we should provide 5 or 6 reading passages, followed by 50 multiple choice items testing whether the candidate has understood the passage. There is no vocabulary-list we have to stick to. We do not work according to a matrix or hierarchy, but formulate questions always bearing in mind: is this the information the writer of the article would also call relevant? Is this what he wanted to point out to us? This method, though not very scientific or exact, produces question types that could be fit into a matrix, in other words: the questions do seem to appeal to various subskills of the reading skill. So there are questions asking the candidates to reproduce, draw a conclusion, recognize the tone or style, state the main point. All in MC, of course, so the candidate has to recognize a good reproduction, conclusion, etc. The question is how to prove that the test is valid. We have tried to do some research in various ways.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. van Weeren

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL016709

BEST COPY AVAILABLE

First we tried factor analysis. We factor-analyzed the results on subskill-tests and the traditional multiple choice test. This confirmed the hypothesis that the traditional test appeals to a considerable number of subskills of reading comprehension.

In the second stage we compared the traditional reading comprehension test to two other forms of multiple choice reading comprehension test, this time by means of Rasch analysis. I shall not go into the results of the second stage. It made us only conclude that the trait or skill - the reading skill - did not differ in the three tests. We could not say that any test was better than or preferable to the other two tests. The analysed data in the third stage however, gave us a little more food for thought.

In the third stage we decided we would like to be able to confirm or deny the repeated criticism of teachers that their pupils cannot really tell what the passages in the test are mainly about, even though their scores on the multiple choice test are satisfactory. In other words, candidates do know how to choose the right answer in multiple choice items, but they would not be sufficiently able to formulate in their own words the right answer to an open-ended question, and, above all, they would not be able to summarize the passages in the test.

So what we did was: we chose two passages from a traditional multiple choice examination test of English reading comprehension. Together with the original items this would be test number one. Test number two consisted of the same two passages, this time followed by open-ended questions. The questions were more or less the same as the stems of the multiple choice items in test 1. Test three consisted of again the same two passages, which the candidates had to summarize in their own words, and in Dutch. There was a maximum number of words. We were of course aware that tests 2 and 3, with their 'open' nature would cause problems. They would have to be rated in such a way that each candidate would get a fair treatment. We spent a lot of energy on the response models, the way of scoring the answers, pretesting etc.

I have now come to the method we used to analyse the data we obtained after the results of the tests had been returned to us. You may already have guessed: Rasch's latent-trait analysis. It permitted us to join the three tests in one model. This would enable us to see whether the tests fitted the model, in other words, we could find out whether the tests measured the same skill. Moreover, the difficulty parameter would show whether the three types of tests differed in difficulty. So with this method we could ask ourselves the questions:

- do the tests fit the model? or perhaps which test fits the model best?
- what are the differences in the difficulty parameters?

Now I would like you to have a look at the handouts. If you look up page 43 in the appendix, you will see the fit analysis. The three tests are considered to be one test of 44 items: the first column on page number one shows the range of scores, split up in 6 scoregroups. Group one has

the lowest scores, group 6 the highest. So the first column deals with item difficulty; it gives it the item characteristic curve. The second column, again consisting of the 6 groups, shows us the number of unexpected answers. For every item and every group the difference between what has been observed and what should have been observed is calculated. The difference is made visible by means of stating the number of too many or too few right answers.

The third column gives the model fit, per group of scores and per item.
The last column shows the model fit per item, after all the groups have been taken together. It shows what is called the fit mean square. Looking at the fit mean square is the quickest way to find out which items do not fit into the model, or to be more positive, to see if the test is homogeneous. Before telling you our conclusions from these data I should point out four causes that could lead to lack of fit. First, heterogenity of items. Speaking in terms of factorial itemstructure this would mean, that more than one factor, one variable, is necessary to describe the test items.

Secondly, lack of so called local stochastic independence could lead to lack of fit. This may occur when a variable, not permitted by the Rasch model plays a certain role. This happens for instance when a candidate's memory helps him to find the right answer.

A third cause may lie in persons guessing the right answer. The Rasch model does not allow guessing, but in reality, in a testing situation, it is very difficult to eradicate guessing. When guessing has been frequent, one will find in the lowest scoregroups an overrepresentation of right answers with the difficult items and vice versa an underrepresentation with the easy items. Of course in the case of open-ended questions and a summary guessing cannot play a role.

A fourth cause could be 'item bias'. This occurs when certain items have systematically been too easy or too difficult for one or more identifyable subgroup within the sample.

In order to explain lack of fit, one should first think of lack of local stochastic independence, guessing, and item bias. If these cannot explain the items in question, then item-heterogenity is to blame.

Now, do we in fact have to think of causes of lack of fit in this particular case? Are there many items with a bad fit? Not too many. I have indicated on the analysis which items formally do not fit: numbers 6, 9, 14 and 43.

So our most important general conclusion from this Rasch analysis is that there is one trait, measured by all three different reading comprehension tests. The three tests deal with the same skill. This is of course an agreeable conclusion to come to, if one, for various reasons, would like to stick to the traditional multiple choice test in the final examinations. But if one looks

at the data in detail, there is some more to be said. I'll come back to this later. First I would like to discuss shortly the 4 items that had bad fit.

What could have caused them to be ill-fitting? It is remarkable by the way, that 3 of these 4 items are in the multiple choice test.

A close look at these 3 items shows that there are flaws within the items. They are either too confusing, adding the variable: which I call puzzle(-mindedness), or they require the Dutch pupil to understand typically English ways of expressing feeling or attitude such as when an Englishman uses an understatement. So in the first case we think the bad fit may have been caused by guessing: the candidates just did not know what to choose. The lack of fit in the item which required an understanding of the English way of expressing feeling or attitude may be due to heterogenity as there may be an additional trait involved: a trait having to do with the personality of the candidate being tested.

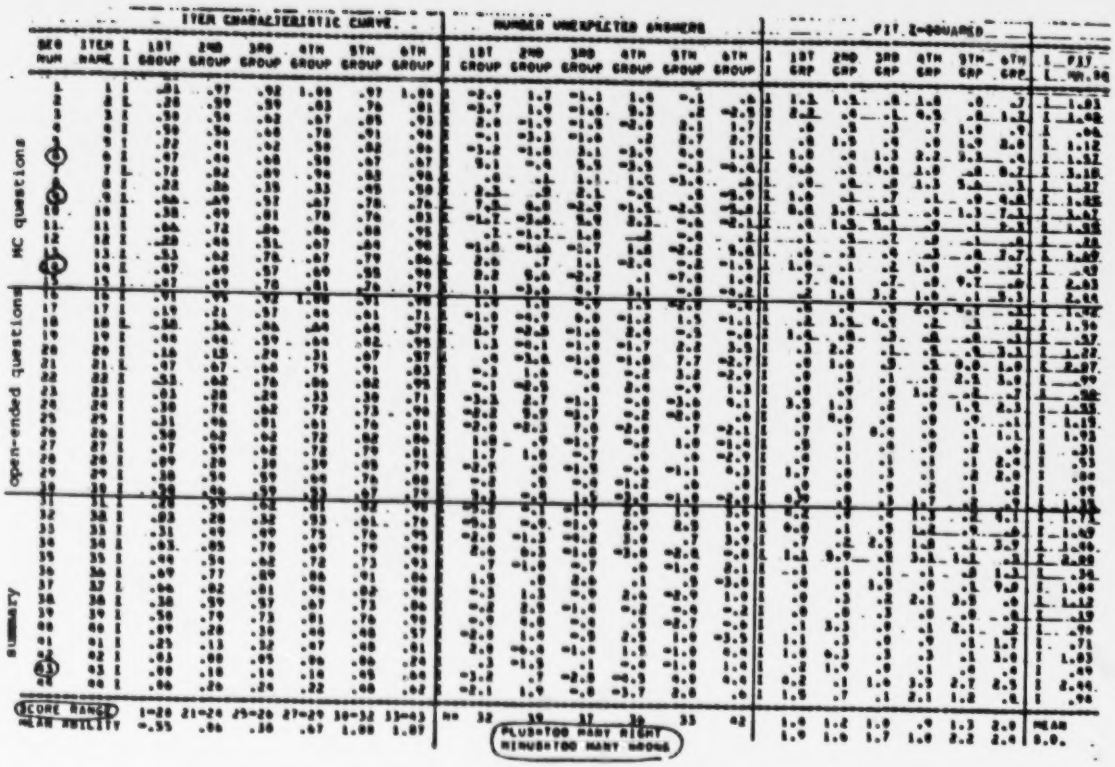
The fourth item, number 43, in the summary, may simply have been too difficult, requiring too much knowledge of English of the candidates concerned. It is interesting to note that knowledge of English seems to become a separate variable, once a certain level of difficulty has been passed. Before passing that level it might be reading comprehension that matters, and not reading in a foreign language that matters.

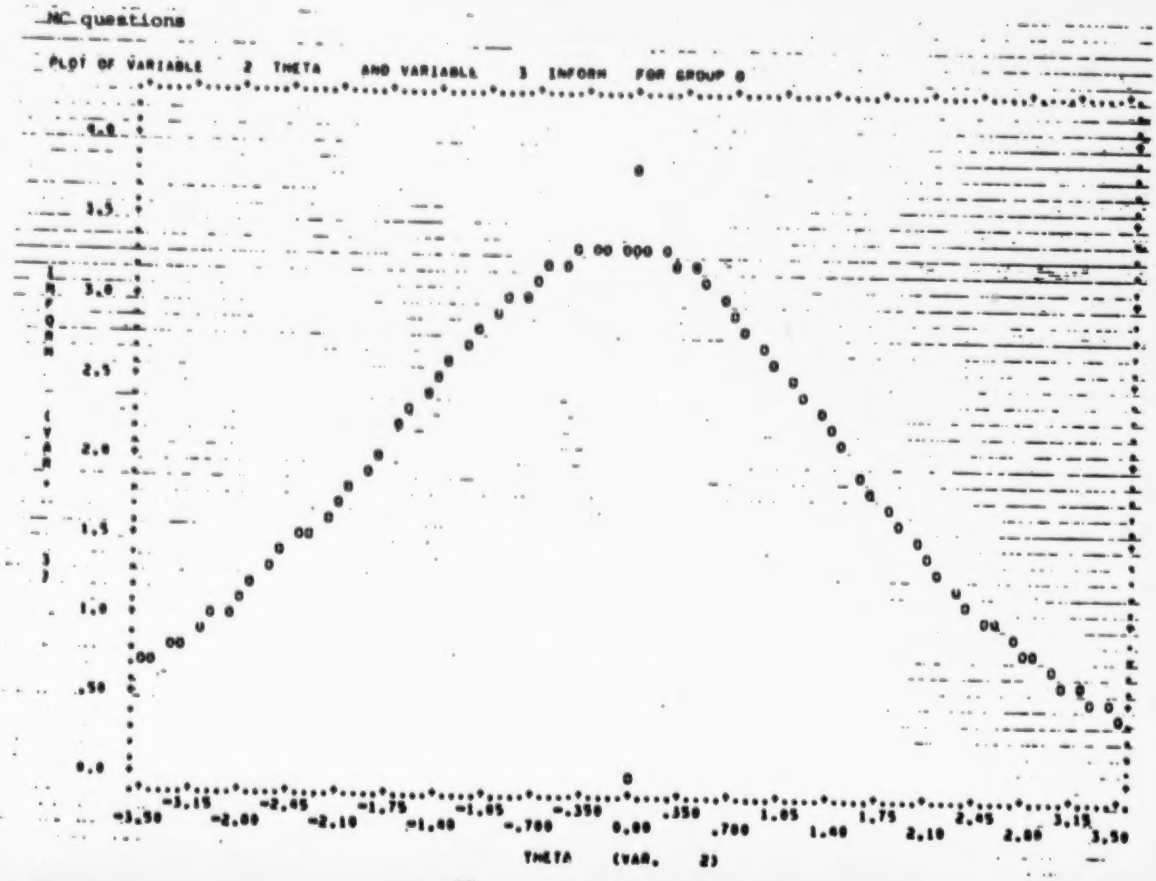
All in all, there are only four items deviating from the test do fit the model. Our first question was answered. You may remember that we also wanted to have a look at the difficulty parameters. They show that the tests do differ in difficulty. The multiple choice test has been the easy one, followed subsequently by the open-ended questions and the summary. In other words: the more open the test is, the more difficult it is. So in a way the criticism, that pupils can answer multiple choice questions correctly, but are not able to formulate answers or give a summary, is understandable. But it is just a question of difference of difficulty of the tests, not of skill of the candidates.

Is its level of difficulty one more reason to stick to the traditional multiple choice test? I suppose it is not, or should not be. Moreover, I have not yet shown you the information curves of the three tests. We need to look at them too in order to find out the accuracy with which the test measures. You can find the curves on page 44. The ideal situation would be if I could also show to you the so called true ability distribution but certain circumstances have made this impossible. The horizontal axis shows ability, the vertical axis shows information. The majority of the candidates is to be found in the area from .350 in each graphic. Comparison of the three curves shows that compared to the i.c. of the MC test the information curve of the summary descends more slowly. This means that of the two the summary measures most accurately, and especially where candidates with the highest ability are concerned. This might be one point in

favour of the summary, if the purpose of the test is maximally accurate measurement in the area where most candidates are to be found. So here the analysis gives us a hint as to what test to prefer, or what direction out research should take. The open-ended questiontest does not do too badly either, of course.

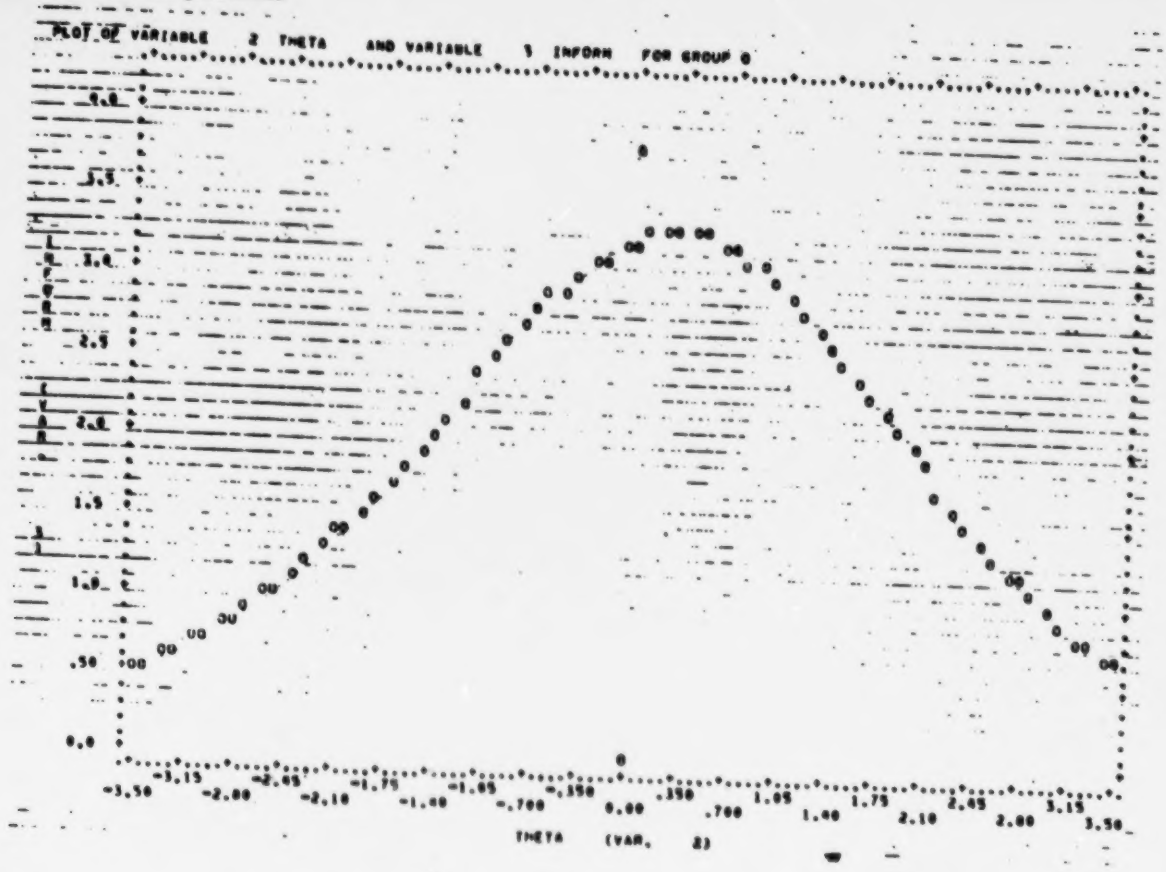
Anyway, we have, for the next step in our research, decided to tackle the summary again. We do know that in the situation of a final examination there are various practical reasons not to choose a summary ever. The construction of the response model and rating, finding teachers who are willing to rate, are all hazardous and time-consuming activities. Even so, we want to have more information on the subject of the summary, as compared to the traditional multiple choice test. This time the passage to be summarized will be another one than in the traditional test. It will be longer, as we feel that this brings the candidate closer to the real life activity. Perhaps it will change the model fit. The passage will be tested in an English and a Dutch version, to find out what part knowledge of English plays.





BEST COPY AVAILABLE

open-ended questions



summary